CONVERGENCE PROPERTIES OF AN ADAPTIVE ALGORITHM FOR LINEAR THRESHOLD ELEMENTS

ROBERT R. PEARSON

CONVERGENCE PROPERTIES OF AN ADAPTIVE ALGORITHM FOR LINEAR THRESHOLD ELEMENTS

by

Robert R. Pearson
Lieutenant, United States Naval Reserve
B.A., University of Connecticut, 1959

[redacted]

Submitted in partial fulfillment
for the degree of

MASTER OF SCIENCE

with major in

MATHEMATICS

from the

UNITED STATES NAVAL POSTGRADUATE SCHOOL
May 1966

ABSTRACT

"Linear threshold element" is the generic term for a device which forms the sum $a_0 + a_1x_1 + a_2x_2 + \ldots + a_dx_d$ from an input vector $(x_1, x_2, \ldots, x_d)$ and yields one of two outputs depending on whether or not the sum is positive. A pattern classification machine may utilize a linear threshold element along with a controller which receives the one of the two values corresponding to correct classification of the input vector. The purpose of the controller is to modify the gain vector $(a_0, a_1, \ldots, a_d)$ so that the next input vector has a greater likelihood of being correctly classified by the threshold element.

This likelihood depends on the value of the gain vector and an adaptive algorithm of the "steepest descent" variety can be used to attempt to adjust the gain vector to its optimal value as the machine is exposed to a stationary sequence of statistically independent input vectors. The components of these vectors are commonly two valued, and it has been shown that convergence of the expected value of the gain vector is dependent on the value of the adjustment parameter, the values of the components, and the distribution of the input vectors. It is shown herein that a bound on the adjustment parameter, simply related to the values of the input components, is sufficient to insure this convergence. The variance of the gain vector is derived under the assumptions of a uniform input sequence and oppositely signed components of equal magnitude and it is shown that a similar bound on the adjustment parameter implies convergence of the variance. The variance is graphed under representative conditions.

TABLE OF CONTENTS

1. Introduction.

"Linear threshold element" (LTE) is the generic term for a device which forms the sum $a_0 + a_1x_1 + a_2x_2 + \ldots + a_dx_d$ from an input vector $(x_1, x_2, x_3, \ldots, x_d)$ and yields one of two outputs depending on whether or not the sum is positive. The components of the input vectors are commonly two valued, and therefore the total number of possible input vectors is 2^d. When used in a pattern classification machine (PCM) the output of the LTE classifies each input into one of two classifications. This classification may or may not be correct. The correct classification is given by the environment, a fixed but unknown function defined on the same set of input vectors, whose output is either of the two values of the LTE.

As an example of an environment, consider a handwritten letter of the alphabet which is "read" by an array of mark sensing devices. The input to the environment is the pattern from the sensing devices and the output is whether or not the letter which was read is a particular letter. Consider also medical diagnosis. The electrocardiagraph of a particular human heart may be sensed as above and the output of the environment is whether or not a particular anomaly is present. In either case if the same sensing pattern is presented to an LTE, it will also make a classification. The number of similar applications is large.

In addition to the LTE, a PCM may use a controller. This device receives the environmental response to an input vector and attempts to modify the gain vector $(a_0, a_1, a_2, \ldots, a_d)$ so that the next input vector has a greater likelihood of being correctly classified by the LTE. Several algorithms have been proposed to adjust the gain vector.[1] The method under consideration is the steepest descent algorithm of

Widrow and Hoff. [2,3] It will be discussed in section 2.

The gain vector is given an initial setting a(1), and then the PCM is exposed to a potentially infinite sequence of input vectors $\{X(n)\}$ and the corresponding sequence of correct output values from the environment $\{f[X(n)]\}$, $n = 1,2,3,\ldots$. After each input vector is presented, the gain vector is adjusted by the controller to yield the sequence $\{a(n)\}$. Under certain conditions, this sequence will converge to a terminal vector a* which is optimal in some sense for the correct classification of an input vector by the LTE.

It is assumed that the sequence of input vectors is a strictly stationary stochastic sequence of independent random variables, ie. $X(n) = X$, where X is a random variable whose statistical properties are completely described by the probability vector $P = (p_1, p_2, \ldots, p_{2^d})$, and $p_j = \Pr\{X = X^j\} > 0$ for $j = 1, 2, 3, \ldots, 2^d$ and $\sum_{j=1}^{2^d} p_j = 1$. A frequent example is the uniform input sequence: $p_j = 2^{-d}$ for all j, which implies that the occurrence of each of the 2^d input vectors is equally likely.

The LTE is used to predict the environment, and a measure of its ability to perform this task is the state of the PCM, S(a), defined as the expected value of the squared difference between the responses of the LTE and the environment with respect to the random variable X. The state of the PCM is zero if and only if the responses of the LTE and the environment are equal for all input vectors. The task of the controller is to minimize S with respect to the gain vector.

Due to the discontinuity of the step function, the minimization of S will not be without difficulty. Consider an auxilliary measure of the performance of the PCM, Q(a), the expected value of the squared differ-

ence between the response of the environment and the sum $a_0 + a_1x_1 + \ldots + a_dx_d$ which is internally generated by the LTE. This auxilliary measure is the one chosen for minimization and it is believed that the following theorem is correct, although a satisfactory proof is not known to exist.

Theorem 1. If $Q(a^*) \leq Q(a)$ for all gain vectors a,

then $S(a^*) \leq S(a)$ for all gain vectors a.

To summarize the assumptions and notation, let the possible values of the components of the input vectors be q and r, then the LTE, g, is defined on the set

$$B^d = \{ X^j : X^j = (x_0, x_1, \ldots, x_d)^T; \; x_0 = \max[|q|, |r|]; \; x_i \in \{ q, r\} , \; i = 1, 2, \ldots, d; \; j = 1, 2, \ldots, 2^d \} .$$

Let $A^d = \{a: a = (a_0, a_1, \ldots, a_d)^T \}$ be the set of gain vectors. Then $g(X) = \text{sgn}(a^TX)$, where $\text{sgn}(t) = \begin{cases} 1, & \text{if } t > 0 \\ -1, & \text{otherwise,} \end{cases}$

and $R = \{ 1, -1\}$ is the range set of g. The environment, f, also maps B^d into R. Note that each environment could be interpreted as one of the 2^{2^d} Boolean functions of d binary variables. The measures are as follows,

$$S(a) = \overline{[f(X) - g(X)]^2} \quad \text{and}$$

$$Q(a) = \overline{[f(X) - a^TX]^2} .$$

X is a random variable which takes on values in B^d, all with a positive probability, in accordance with the probability vector P. Note that all functions of X are therefore random variables.

2. The Adaptive Algorithm

The problem is to minimize

$$Q(a) = \overline{[\, f(X) - a^T X]^2} = \overline{f^2(X)} - 2\sum_{\ell=0}^{d} a_\ell \overline{f(X)x_\ell}$$

$$+ \sum_{\ell=0}^{d}\sum_{k=0}^{d} a_\ell a_k \overline{x_\ell x_k}.$$

Setting $\frac{\partial Q(a)}{\partial a_i} = 0, \forall i$, yields

$$\overline{f(X)x_i} = \sum_{\ell=0}^{d} a_\ell \overline{x_\ell x_i}\ , \forall i. \tag{1}$$

That value of the gain vector which satisfies equation (1) is the vector a*, which minimizes Q(a). But since f is unknown, the equation cannot be solved directly.

Using the method of steepest descent, the gain vector is modified after each presentation of an input vector:

$$a(n+1) = a(n) + \eta\, \mathrm{grad}Q_n,$$

where $a(n)$ is the value of the gain vector at the time of the n^{th} presentation;

$Q_n = \{ f[X(n)] - a^T(n)X(n) \}^2$;

$X(n)$ is the n^{th} input vector;

$\mathrm{grad}Q_n = -\left(\frac{\partial Q_n}{\partial a_0}, \frac{\partial Q_n}{\partial a_1}, \ldots, \frac{\partial Q_n}{\partial a_d}\right)$;

η is a positive constant.

Each component is adjusted in the direction of decreasing Q_n. Specifically,

$$a_i(n+1) = a_i(n) + 2\eta x_i(n) \{ f[X(n)] - a^T(n)X(n) \}\ , \forall i. \tag{2}$$

It is to be noted that a(n) is a random variable. Now show that $\overline{a(n)}$ converges to a*. Martinez shows this as follows:[4]

Rewrite equation (2) as

$$a(n+1) = a(n) + \beta[\, b_n - C_n\, a(n)] \tag{3}$$

where the adjustment parameter $\beta = 2\eta$, $b_n = f[\, X(n)]X(n)$, and $C_n = X(n)X^T(n)$. Note that $C_n = C_n^T$.

Hence $a(n+1) = \beta\, b_n + [\, I - \beta\, C_n\,]\, a(n)$

$$= d_n + E_n a(n), \tag{4}$$

where $d_n = \beta b_n$, and $E_n = I - \beta\, C_n$.

Expanding recursively and taking expected values,

$$\overline{a(n)} = \overline{d_{n-1}} + \overline{E_{n-1}d_{n-2}} + \overline{E_{n-1}E_{n-2}d_{n-3}} + \cdots$$

$$+ \overline{E_{n-1}E_{n-2}\cdots E_1 a(1)}. \tag{5}$$

Due to the assumptions of independence and stationarity on the input sequence, equation (5) becomes

$$\overline{a(n)} = [\, I + E + E^2 + \ldots + E^{n-2}\,]\, D + E^{n-1}\overline{a(1)},$$

where $E = \overline{E_n}$ and $D = \overline{d_n}$,

$$= [\, I - E]^{-1}\, [I - E^{n-1}]\, D + E^{n-1}\overline{a(1)}. \tag{6}$$

If $\lim_{n\to\infty} E^n = 0$, then

$$\lim_{n\to\infty} \overline{a(n)} = [\, I - E\,]^{-1} D.$$

It is then shown that if $C = \overline{C_n}$ is positive definite, and if $\beta < \frac{2}{\lambda}$, where λ is the largest eigenvalue of C, then $\lim_{n\to\infty} E^n = 0$. Hence, if C is positive definite, and the positive constant η is less than

the reciprocal of the largest eigenvalue of C, then the modification given by equation (2) will cause $\overline{a(n)}$ to converge to a*, since $[\, I - E \,]^{-1} D = C^{-1} b$, where $b = \overline{b_n}$, and if a' is the terminal value of $\overline{a(n)}$, then $b = Ca'$, which is the minimizing equation (1).

Under what conditions is C positive definite? Consider $C^j = X^j X^{jT}$, and let Z be a non zero vector. Then $Z^T C^j Z = Z^T X^j X^{jT} Z = (Z^T X^j)^2 \geq 0$, which shows that C^j is positive semi definite for all j.

$C = \overline{C_n} = \sum_{j=1}^{2^d} p_j C^j$, and if Z is as above, then

$$Z^T C Z = \sum_{j=1}^{2^d} p_j Z^T C^j Z \geq 0.$$

Therefore C is always positive semi definite. That it is, in fact, always positive definite is shown by contradiction. Assume $\exists\ W \ni$ $W^T C W = 0$.

Then $p_j W^T C^j W = 0$, for all j. Since $p_j > 0$ for all j,

$W^T C^j W = 0$ for all j, which implies that

$(W^T X^j)^2 = 0$ for all j, and

$$W^T X^j = 0 \text{ for all } j. \tag{7}$$

Equation (7) is a linear system of 2^d equations, and if the vector W is a solution, then it must satisfy a subsystem of d+1 of the equations. Let $\{X^{b_1}, X^{b_2}, \ldots, X^{b_{d+1}}\}$ be a set of linearly independent vectors from B^d. Such a set exists since B^d spans d+1 space. Hence

$$W^T (X^{b_1} X^{b_2} \ldots . X^{b_{d+1}}) = 0. \tag{8}$$

Now a nontrivial solution to (8) exists if and only if

$$| X^{b_1} X^{b_2} \ldots . X^{b_{d+1}} | = 0.$$

But this set of vectors is linearly independent and their determinant is non zero. Hence for all non zero vectors W, $W^T CW \neq 0$, and C is always positive definite.

What is the range of the eigenvalues of C? For each C^j, $\text{diag}(C^j) = ((x_o^j)^2, (x_1^j)^2, \ldots, (x_d^j)^2)$, and it follows that $\text{diag}(C) = (e_0, e_1, \ldots, e_d)$, where

$$e_i = \sum_{j=1}^{2^d} p_j (x_i^j)^2 \text{ for all } i.$$

There is no loss of generality in assuming $|q| \leq |r|$, and in that case $e_0 = r^2$ and

$$q^2 < e_i < r^2 \text{ for } i = 1, 2, \ldots, d.$$

Hence $\text{trace } (G) = \sum_{i=0}^{d} e_i < (d+1)r^2$.

Let $\lambda_0, \lambda_1, \ldots, \lambda_d$ be the eigenvalues of C. Then, since $\text{trace } (G) = \sum_{i=0}^{d} \lambda_i$ and $\lambda_i > 0$ for all i,

$$\lambda < \sum_{i=0}^{d} \lambda_i < (d+1)r^2.$$

It then follows that if the adjustment parameter $\beta \leq \frac{2}{(d+1)r^2}$, then $\overline{a(n)}$ will converge to a*, since C is always positve definite and

$$\beta \leq \frac{2}{(d+1)r^2} \quad \Rightarrow \quad \beta < \frac{2}{\lambda} .$$

The remainder of this section examines the convergence of $\overline{a(n)}$ under the following assumptions.

(1) The components of the input vectors are of equal magnitude and oppositely signed, ie. $0 < -q = r$.

(2) The input sequence is uniform, ie. $p_j = 2^{-d}$ for all j.

With these assumptions C will be shown to reduce to $r^2 I$, which simplifies further analysis without sacrificing a great deal of applicability, since any PCM can be reworked to make the transformation of (1) above, and (2) above is a common ad hoc approach to a practical situation.

Consider the following constructive scheme for the input vectors which is analogous to the binary representation of the integers.

$$X^1 = (r, -r, -r, \ldots, -r, -r, -r)^T$$
$$X^2 = (r, -r, -r, \ldots, -r, -r, r)^T$$
$$X^3 = (r, -r, -r, \ldots, -r, r, -r)^T$$
$$X^4 = (r, -r, -r, \ldots, -r, r, r)^T$$
$$\vdots \qquad\qquad \vdots$$
$$X^{2^d} = (r, r, r, \ldots, r, r, r)^T$$

This scheme could also be displayed as the matrix $M = (m_{ij})$ with each element of the form $\frac{x_i^j}{r}$ where (contrary to the usual convention) i is the column index - which refers to the components of a particular input vector - and j is the row index - which refers to a particular vector from B^d. The order of M is $2^d(d + 1)$. An analytic expression for m_{ij} is

$$\begin{cases} -1, & \text{if } (j-1) \bmod 2^{d-i+1} < 2^{d-1} \\ 1, & \text{otherwise.} \end{cases}$$

$$
M = \begin{array}{c}
\begin{array}{cccccccc} \text{col } 0 & \text{col } 1 & \text{col } 2 & & \text{col } d-3 & \text{col } d-2 & \text{col } d-1 & \text{col } d \end{array} \\
\begin{pmatrix}
1 & -1 & -1 & \cdots & -1 & -1 & -1 & -1 \\
1 & -1 & -1 & \cdots & -1 & -1 & -1 & 1 \\
1 & -1 & -1 & \cdots & -1 & -1 & 1 & -1 \\
1 & -1 & -1 & \cdots & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & \cdots & -1 & -1 & -1 & -1 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\
1 & -1 & -1 & \cdots & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & \cdots & -1 & -1 & -1 & -1 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\
1 & -1 & 1 & \cdots & 1 & 1 & 1 & 1 \\
1 & 1 & -1 & \cdots & -1 & -1 & -1 & -1 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\
1 & 1 & 1 & \cdots & 1 & 1 & 1 & 1
\end{pmatrix}
\end{array}
\begin{array}{l}
\text{row } 2^0 \\
\text{row } 2^1 \\
\text{row } 2^1 + 1 \\
\text{row } 2^2 \\
\text{row } 2^2 + 1 \\
\\
\text{row } 2^{d-2} \\
\text{row } 2^{d-2} + 1 \\
\\
\text{row } 2^{d-1} \\
\text{row } 2^{d-1} + 1 \\
\\
\text{row } 2^d
\end{array}
$$

Now it is necessary to show that the column vectors of M are mutually orthogonal. Using the method of induction of the dimension of the PCM, let d = 1. Then

$M_1 = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = (N_0 \; N_1)$ where N_0 is the column vector $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and and N_1 is the column vector $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$. There is only one pair of vectors to check. $N_0^T N_1 = -1 + 1 = 0$. The vectors are orthogonal. Now let d = ℓ

and $M_\ell = (C_0C_1C_2 \cdots C_\ell)$ where C_i is the $i^{\underline{th}}$ column vector of M_ℓ. The induction assumption is that the column vectors of M_ℓ are mutually orthogonal. Then if K_i is the $i^{\underline{th}}$ column vector for the dimension $\ell + 1$,

$$M_{\ell+1} = (K_0K_1K_2\cdots K_{\ell+1}) \text{ which can be partitioned}$$

with respect to rows as

$$\begin{pmatrix} C_0 - C_0C_1C_2 \cdots C_\ell \\ C_0 \quad C_0C_1C_2 \cdots C_\ell \end{pmatrix}.$$

With this partitioning any column vector of $M_{\ell+1}$ can be expressed as a direct sum (a physical concatenation) of the column vectors of M_ℓ, to wit:

[5] $K_0 = C_0 \dot{+} C_0$, $K_1 = (-C_0) \dot{+} C_0$, and $K_i = C_{i-1} \dot{+} C_{i-1}$ for $i = 2, 3, \ldots, \ell+1$. Then any product of the form $K_0^TK_i$ for $i = 2, 3, \ldots, \ell+1$ can be expressed as

$$(C_0 \dot{+} C_0)^T(C_{i-1} \dot{+} C_{i-1}) = C_0^TC_{i-1} + C_0^TC_{i-1} = 0 + 0 = 0,$$

with $K_0^TK_1 = (C_0 \dot{+} C_0)^T((-C_0) \dot{+} C_0) = -C_0^TC_0 + C_0^TC_0 = 0.$

Similarly, any product of the form $K_1^TK_i$ for $i = 2, 3, \ldots, \ell+1$ can be expressed as $((-C_0) \dot{+} C_0)^T(C_{i-1} \dot{+} C_{i-1})$

$= -C_0^TC_{i-1} + C_0^TC_{i-1} = 0$, and any product of the form $K_i^TK_j$ for $1 < i < j \leq \ell+1$ can be expressed as

$$(C_{i-1} \dot{+} C_{i-1})^T(C_{j-1} \dot{+} C_{j-1})$$

$$= C_{i-1}^TC_{j-1} + C_{i-1}^TC_{j-1}$$

$$= 2C_{i-1}^TC_{j-1}$$

$= 0$, since all C_i's are orthogonal.

Thus the column vectors of $M_{\ell+1}$ are mutually orthogonal, and by induction it is true that for all positive integral values of d, the columns of the matrix M are mutually orthogonal.

Now consider the elements of $C = \sum_{k=1}^{2^d} p_k C^k$ where $p_k = 2^{-d}$, and the input components are $\pm r$.

$$c_{ij} = 2^{-d} \sum_{k=1}^{2^d} c_{ij}^k = 2^{-d} \sum_{k=1}^{2^d} (X^k X^{kT})_{ij}$$

$$= 2^{-d} \sum_{k=1}^{2^d} x_i^k x_j^k$$

$$= 2^{-d} r^2 \sum_{k=1}^{2^d} m_{ik} m_{jk}, \text{ since } m_{ij} = \frac{x_i^j}{r} .$$

This reduces to $c_{ij} = r^2 \delta_{ij}$, where δ_{ij} is the Kronecker delta,

since $\sum_{k=1}^{2^d} m_{ik} m_{jk} = 0$ if $i \neq j$, due to the orthogonality of the column vectors of the matrix M. Hence $C = r^2 I$.

Under these assumptions equation (6) becomes

$$\overline{a(n)} = [I - E]^{-1} [I - E^{n-1}] D + E^{n-1} a(1)$$

$$= [\beta C]^{-1} [I - (I - \beta C)^{n-1}] D + [I - \beta C]^{n-1} a(1)$$

$$= [\beta r^2 I]^{-1} [I - (I - \beta r^2 I)^{n-1}] D + [I - \beta r^2 I]^{n-1} a(1)$$

$$= \frac{1}{\beta r^2} [1 - (1 - \beta r^2)^{n-1}] \beta b + (1 - \beta r^2)^{n-1} a(1)$$

$$= \frac{1}{r^2} [1 - (1 - \beta r^2)^{n-1}] Ca^* + (1 - \beta r^2)^{n-1} a(1)$$

$$= \left[1 - \left(1 - \beta r^2\right)^{n-1}\right] a^* + \left(1 - \beta r^2\right)^{n-1} a(1) \tag{9}$$

$$= a^* - \left[a^* - a(1)\right]\left(1 - \beta r^2\right)^{n-1} . \tag{10}$$

Equation (1) can be written as $b = Ca^*$, which becomes $b = r^2 a^*$, and the optimal value of the gain vector, $a^* = \frac{b}{r^2}$. Then

$$Q(a^*) = \overline{f^2(X)} - 2\sum_{\ell=0}^{d} a_\ell^* \overline{f(X) x_\ell} + \sum_{\ell=0}^{d} \sum_{k=0}^{d} a_\ell^* a_k^* \overline{x_\ell x_k}$$

$$= 1 - 2\sum_{\ell=0}^{d} a_\ell^* r^2 a_\ell^* + \sum_{\ell=0}^{d} \sum_{k=0}^{d} a_\ell^* a_k^* r^2 \delta_{\ell k}$$

$$= 1 - r^2 (a^*)^2 \tag{11}$$

From equation (11) it is evident that if $Q(a^*)$, the optimum of the minimization effort, is near zero, then $(a^*)^2$ is near r^{-2}. Note also the range of $(a^*)^2$, between zero and r^{-2}.

3. The Variance of the Gain Vector.

Consider the variance of a(n), $V[a(n)] = \overline{a^2(n)} - \overline{a(n)}^2$. It is necessary to compute $\overline{a^2(n)}$, which will be done under assumption (1), the components are $\pm r$. From equation (4),

$$a(n+1) = d_n + E_n a(n),$$

$$\text{where } d_n = \beta b_n = \beta f[X(n)]\, X(n),$$

$$E_n = I - \beta C = \beta X(n) X^T(n),$$

$$\text{and let } 0 < \beta \leq \frac{2}{(d+1)r^2}.$$

Then $$a^2(n+1) = [d_n^T + a^T(n)E_n^T][d_n + E_n a(n)]$$

$$= d_n^2 + 2d_n^T E_n a(n) + a^T(n)E_n^T E_n a(n),$$

which can be written as

$$a^2(n+1) = \beta^2 r^2(d+1) + 2[1 - \beta r^2(d+1)]\, d_n^T a(n)$$

$$+ a^T(n)\{I + \beta[\beta r^2(d+1) - 2]\, C_n\}\, a(n), \qquad (12)$$

since $d_n^2 = \beta b_n^T \beta b_n = \beta^2 f[X(n)]\, X^T(n) f[X(n)]\, X(n)$

$$= \beta^2 X^T(n) X(n)$$

$$= \beta^2 r^2(d+1),$$

and $d_n^T E_n = d_n^T [I - \beta C_n] = \beta f[X(n)]\, X^T(n)\, [I - \beta X(n) X^T(n)]$

$$= \beta f[X(n)][X^T(n) - \beta r^2(d+1) X^T(n)]$$

$$= [\, 1 - \beta r^2(d+1) \,]\, d_n^T,$$

and $$E_n^T E_n = I - 2\beta C_n + \beta^2 C_n^T C_n$$

$$= I - 2\beta C_n + \beta^2 r^2 (d+1) C_n.$$

$$= I + \beta[\, \beta r^2(d+1) - 2 \,] C_n.$$

Taking the expected value of equation (12),

$$\overline{a^2(n+1)} = \beta^2 r^2 (d+1) + 2[\, 1 - \beta r^2(d+1)] \, \overline{d_n^T a(n)}$$

$$+ \overline{a^2(n)} + \beta[\beta r^2(d+1) - 2] \, \overline{a^T(n) C_n a(n)}. \qquad (13)$$

Before considering the expected values of $d_n^T a(n)$ and $a^T(n) C_n a(n)$, recall the following theorem, as stated by Halmos. [6]

If $\{ f_{ij} : i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, n_i \}$ is a set of independent functions, if φ_i is a real valued, Borel measurable function of n_i real variables, $i = 1, 2, \ldots, k$, and if $f_i(x) = \varphi_i(f_{i1}(x), \ldots, f_{in_i}(x))$, then the functions $f_1, \ldots, f_k$ are independent.

Now consider the set of independent random variables $\{X(1), X(2), \ldots, X(n-1), X(n)\}$. By definition $d_n = \beta f[\, X(n)] \, X(n)$ and is defined on the subset $\{X(n)\}$. On the other hand a(n) is defined on the complementary subset $\{ X(1), \ldots, X(n-1)\}$ by the recursion relation of equation (2). Therefore, applying the theorem to each component of these vectors, the conclusion is that a(n) and d_n are independent random variables. This being the case, the expected value of $d_n^T a(n)$ is the product of the expected values of d_n and a(n). Hence,

$\overline{d_n^T a(n)} = \overline{d_n^T a(n)} = \beta \, b^T \overline{a(n)}$. Then using $b = Ca^*$ and equation

(6),

$$\overline{d_n^T a(n)} = \beta (a^*)^T C \{ (I - E)^{-1} (I - E^{n-1}) D + E^{n-1} a \}$$

where $a = a(1)$,

$$= \beta (a^*)^T \{ [I - (I - \beta C)^{n-1}] a^* + (I - \beta C)^{n-1} a$$

$$= \beta (a^*)^T C a^* - \beta (a^*)^T C (I - \beta C)^{n-1} (a^* - a) \qquad (14)$$

Next consider the expected value of $a^T(n) C_n a(n)$.

$$\overline{a^T(n) C_n a(n)} = \overline{\sum_{i=0}^{d} a_i(n) \sum_{j=0}^{d} x_i(n) x_j(n) a_j(n)}$$

$$= \sum_{i=0}^{d} \sum_{j=0}^{d} \overline{a_i(n) a_j(n) x_i(n) x_j(n)} , \qquad (15)$$

since a(n) and X(n) satisfy the hypothesis of the independence theorem above.

The goal now is to express this expected value as $\overline{a^2(n)} G(C)$, where G is some unknown function of C. Equation (15) is close, but it contains terms in $\overline{a_i(n) a_j(n)}$ with $i \neq j$, which have not yielded to analysis. Is it possible that these cross product terms vanish? Or is it possible that their coefficients $\overline{x_i(n) x_j(n)}$ vanish for $i \neq j$? The latter is, in fact, exactly the conclusion arrived at by assumption (2), the uniform input sequence. Under this assumption the derivation can proceed, and it will, justified by the urgent need for <u>some</u> results, however special.

With $C = r^2 I$ used in the expected value terms, equation (5) becomes

$$\overline{a^T(n) C_n a(n)} = \sum_{i=0}^{d} r^2 \overline{a_i^2(n)} = r^2 \overline{a^2(n)},$$

and equation (14) becomes

$$\overline{d_n^T a(n)} = \beta r^2 (a^*)^2 - 2\beta r^2 (a^*)^T (a^* - a)(1 - \beta r^2)^{n-1}.$$

Finally returning to equation (13),

$$\overline{a^2(n+1)} = \beta^2 r^2 (d+1) + 2\beta r^2 [1 - \beta r^2 (d+1)] (a^*)^2$$

$$-2\beta r^2 [1 - \beta r^2 (d+1)] (a^*)^T (a^* - a)(1 - \beta r^2)^{n-1}$$

$$+ [(1 - \beta r^2)^2 + \beta^2 r^4 d] \overline{a^2(n)} \quad . \tag{16}$$

The structure of equation (16) is more readily apparent if the following substitutions are made for the constants of the process.

Let $\sigma = \beta r^2 \{ \beta (d+1) + 2[1 - \beta r^2 (d+1)] (a^*)^2 \}$,

$\gamma = -2\beta r^2 [1 - \beta r^2 (d+1)] (a^*)^T (a^* - a)$,

$\delta = (1 - \beta r^2)^2 + \beta^2 r^4 d$, and

$\rho = (1 - \beta r^2)$.

Then $$\overline{a^2(n+1)} = \sigma + \gamma \rho^{n-1} + \delta \overline{a^2(n)} \ , \tag{17}$$

which when reworked recursively one step becomes,

$$\overline{a^2(n+1)} = \sigma + \gamma \rho^{n-1} + \delta [\sigma + \gamma \rho^{n-2} + \delta \overline{a^2(n-1)}]$$

$$= \sigma (1 + \delta) + \gamma (\rho + \delta) \rho^{n-2} + \delta^2 \overline{a^2(n-1)} \ ,$$

and one more step,

$$\overline{a^2(n+1)} = \sigma (1 + \delta + \delta^2) + \gamma (\rho^2 + \rho\delta + \delta^2) \rho^{n-3} + \delta^3 \overline{a^2(n-2)} \ ,$$

and through all the steps back to a(1), is

$$\overline{a^2(n+1)} = \sigma \sum_{i=0}^{n-1} \delta^i + \gamma \sum_{i=0}^{n-1} \rho^{n-i-1}\delta^i + \delta^n a^2(1). \qquad (18)$$

The geometric series may be put into closed form, yielding

$$\sum_{i=0}^{n-1} \delta^i = \frac{1 - \delta^n}{1 - \delta}, \text{ and } \sum_{i=0}^{n-1} \rho^{n-i-1}\delta^i = \frac{\rho^n - \delta^n}{\rho - \delta}.$$

Now replacing n by n-1, and making the above substitutions in equation (18),

$$\overline{a^2(n)} = \sigma\left[\frac{1 - \delta^{n-1}}{1 - \delta}\right] + \gamma\left[\frac{\rho^{n-1} - \delta^{n-1}}{\rho - \delta}\right] + \delta^{n-1}a^2. \qquad (19)$$

The denominators in equation (19) can be rewritten as follows.

$$1 - \delta = \beta r^2[2 - \beta r^2(d+1)], \text{ and}$$

$$\rho - \delta = \beta r^2[1 - \beta r^2(d+1)].$$

Now making all the substitutions for σ, γ, and ρ, and cancelling the new forms of the denominators,

$$\overline{a^2(n)} = \{\beta(d+1) + 2[1 - \beta r^2(d+1)](a^*)^2\}\left\{\frac{1 - \delta^{n-1}}{2 - \beta r^2(d+1)}\right\}$$

$$- 2(a^*)^T(a^* - a)\{(1 - \beta r^2)^{n-1} - \delta^{n-1}\} + \delta^{n-1}a^2. \qquad (20)$$

Collecting all the terms in δ^{n-1} and $(1 - \beta r^2)^{n-1}$,

$$\overline{a^2(n)} = \frac{\beta(d+1) + 2[1 - \beta r^2(d+1)](a^*)^2}{2 - \beta r^2(d+1)}$$

$$+ \left\{\frac{2(a^* - a)^2 + \beta(d+1)[r^2(2(a^*)^T - a^T)a - 1]}{2 - \beta r^2(d+1)}\right\}\delta^{n-1}$$

$$- 2(a^*)^T(a^* - a)(1 - \beta r^2)^{n-1}. \qquad (21)$$

Equation (10) provides $\overline{a(n)}$ under assumptions (1) and (2) which when squared yields

$$\overline{a(n)}^2 = (a^*)^2 - 2(a^*)^T(a^* - a)(1 - \beta r^2)^{n-1}$$

$$+ (a^* - a)^2[\,(1 - \beta r^2)^2\,]^{n-1} . \qquad (22)$$

Combining equations (21) and (22), the variance of the gain vector is

$$V[\,a(n)\,] = \overline{a^2(n)} - \overline{a(n)}^2$$

$$= \frac{\beta(d+1) + 2[\,1 - \beta r^2(d+1)\,](a^*)^2}{2 - \beta r^2(d+1)}$$

$$+\left\{\frac{2(a^* - a)^2 + \beta(d+1)[\,r^2(2(a^*)^T - a^T)a - 1\,]}{2 - \beta r^2(d+1)}\right\}\delta^{n-1}$$

$$- (a^*)^2 - (a^* - a)^2[\,(1 - \beta r^2)^2\,]^{n-1} . \qquad (23)$$

Combining the constant terms leaves

$$V[\,a(n)\,] = \frac{\beta(d+1)[\,1 - r^2(a^*)^2\,]}{2 - \beta r^2(d+1)}$$

$$+\left\{\frac{2(a^* - a)^2 + \beta(d+1)[\,r^2(2(a^*)^T - a^T)a - 1\,]}{2 - \beta r^2(d+1)}\right\}\delta^{n-1}$$

$$- (a^* - a)^2[\,(1 - \beta r^2)^2\,]^{n-1} ,$$

$$\text{where } \delta = (1 - \beta r^2)^2 + \beta^2 r^4 d . \qquad (24)$$

Now that the variance of the gain vector has been derived, the question is whether or not it converges, and if so, to what, and under what conditions? The bounds on β which imply convergence of the mean are zero and $\frac{2}{r^2}$ (under both assumptions). Hence the bounds on $(1 - \beta r^2)^2$ are zero and one. Therefore $\lim_{n \to \infty}[\,(1 - r^2)^2\,]^{n-1} = 0$. On the other hand δ is a quadratic expression in β which is less than one for

β between zero and $\frac{2}{(d+1)r^2}$ and therefore this lesser bound must be observed in order for the term in δ to vanish, and consequently insure the convergence of the variance of the gain vector. Hence, if both assumptions (1) and (2) are valid, and if $0 < \beta < \frac{2}{(d+1)r^2}$, then both $\overline{a(n)}$ and $V[a(n)]$ will converge.

The mean will converge to a*, that a which minimizes Q(a), and the variance to

$$V = \lim_{n \to \infty} V[a(n)] = \frac{\beta(d+1)[\,1 - r^2(a^*)^2\,]}{2 - \beta r^2(d+1)} . \qquad (25)$$

Now recall the relationship from equation (11) between a* and Q(a*) which enables equation (25) to be written as

$$V = \frac{\beta(d+1)Q(a^*)}{2 - \beta r^2(d+1)}$$

, where the bounds on Q are zero and one and can be interpreted as a measure of the complexity of the environment to which the PCM is exposed. Since the derivative of V with respect to β is

$$\frac{2(d+1)Q(a^*)}{[2 - \beta r^2(d+1)]^2}$$

, which is non negative, and since V vanishes for $\beta = 0$, the smaller β, the smaller V.

Exactly how small V must be in order for the state of the PCM to be minimized is an open question. Intuitively, it is felt that S(a) reaches its minimum before (in the input sequence) Q(a) is minimized, and the answer is probably also the answer to Theorem 1. Further work on this problem is indicated. Also, of course, is the need for generalizing this work to apply to any input sequence. Once these details are in order, it should be possible to obtain an analytic expression for the number (or average number) of trials to achieve a minimum for the state of the PCM.

Other algorithms and configurations which may yield to analysis can be found in Nilsson. [1]

In the appendix are the results of evaluating the variance of the gain vector for some representative values of the parameters.

APPENDIX I

Assumptions: (1) The input components are ± 1.

(2) The input sequence is uniform.

(3) $0 < \beta < \frac{2}{d+1}$

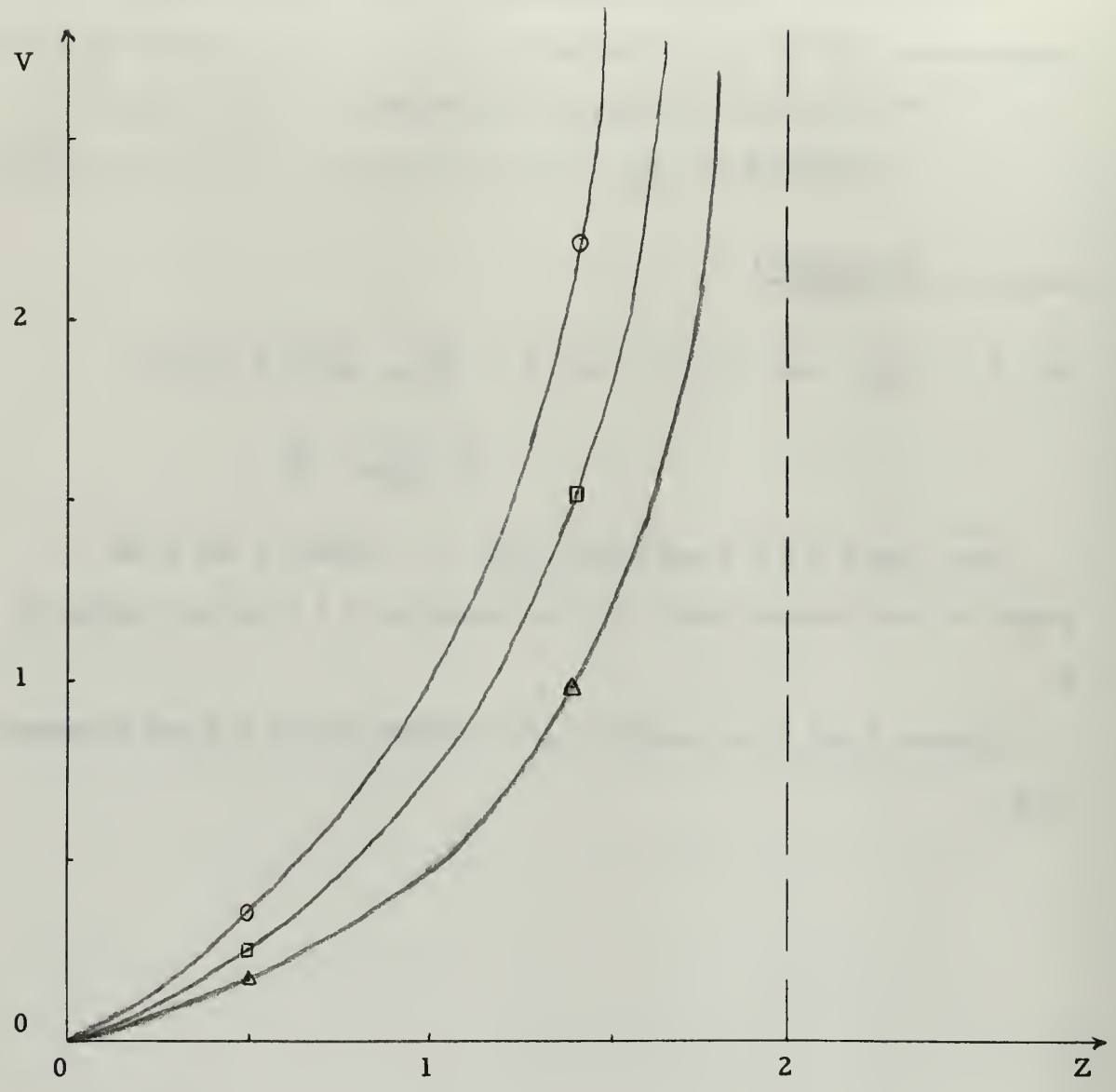
Then $V = \frac{\beta(d+1)Q(a^*)}{2 - \beta(d+1)}$.

Let $\beta = \frac{Z}{d+1}$, then $0 < Z < 2$ and $V = \frac{ZQ}{2-Z}$ where $Q = Q(a^*)$

$$= Q\left[\frac{2}{2-Z} - 1\right].$$

Note that $0 < Q < 1$ and observe Fig. 1. Figures 2 and 3 are graphs of the variance when $Q = 0$ and therefore $V = 0$ for all values of Z.

Figures 4 and 5 are graphs of the variance when $Q = \frac{3}{4}$ and V depends on Z.



○ Q = 1.00

□ Q = .75

△ Q = .50

Figure 1. $V = Q[\frac{2}{2-Z} - 1]$

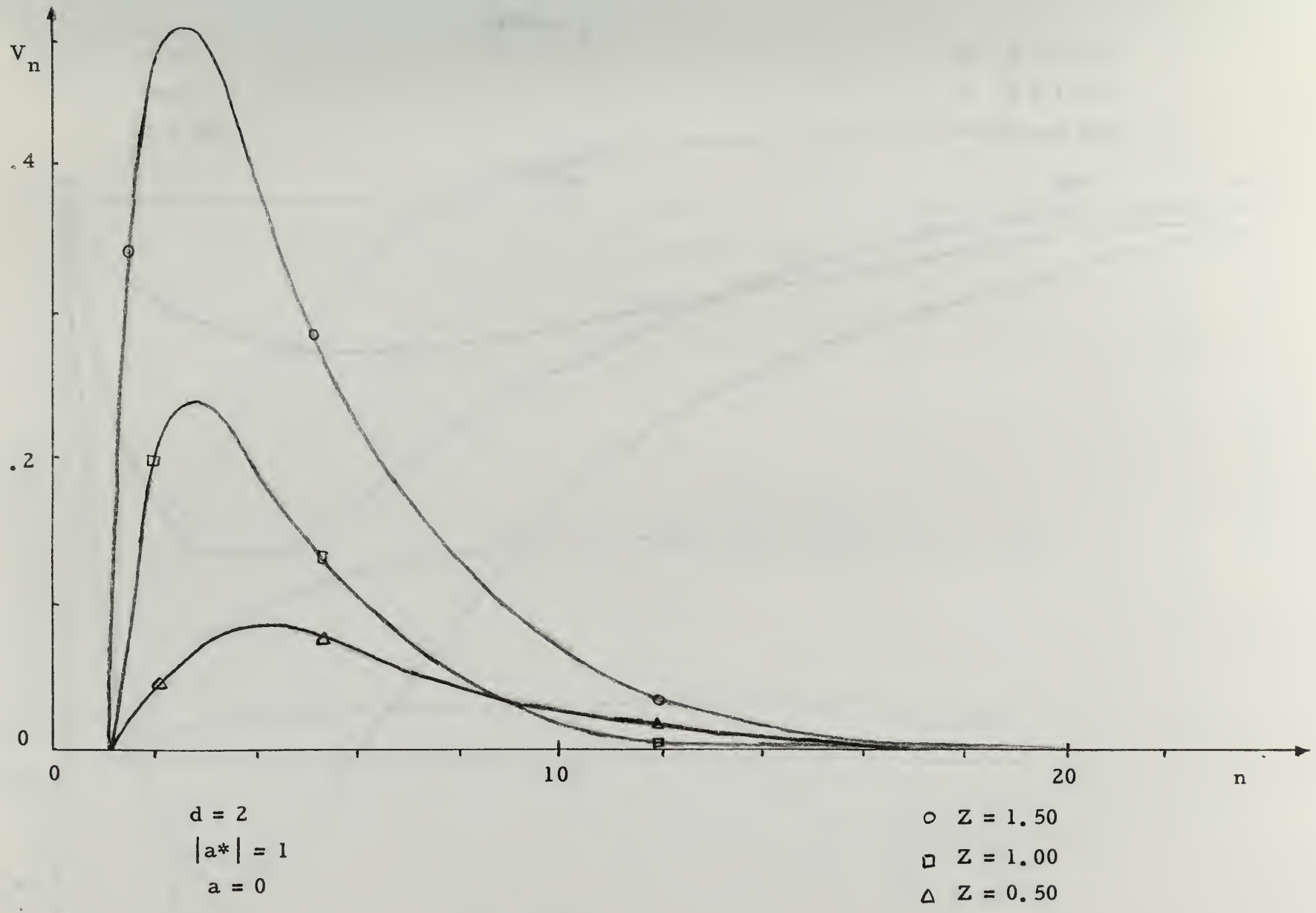


Figure 2.

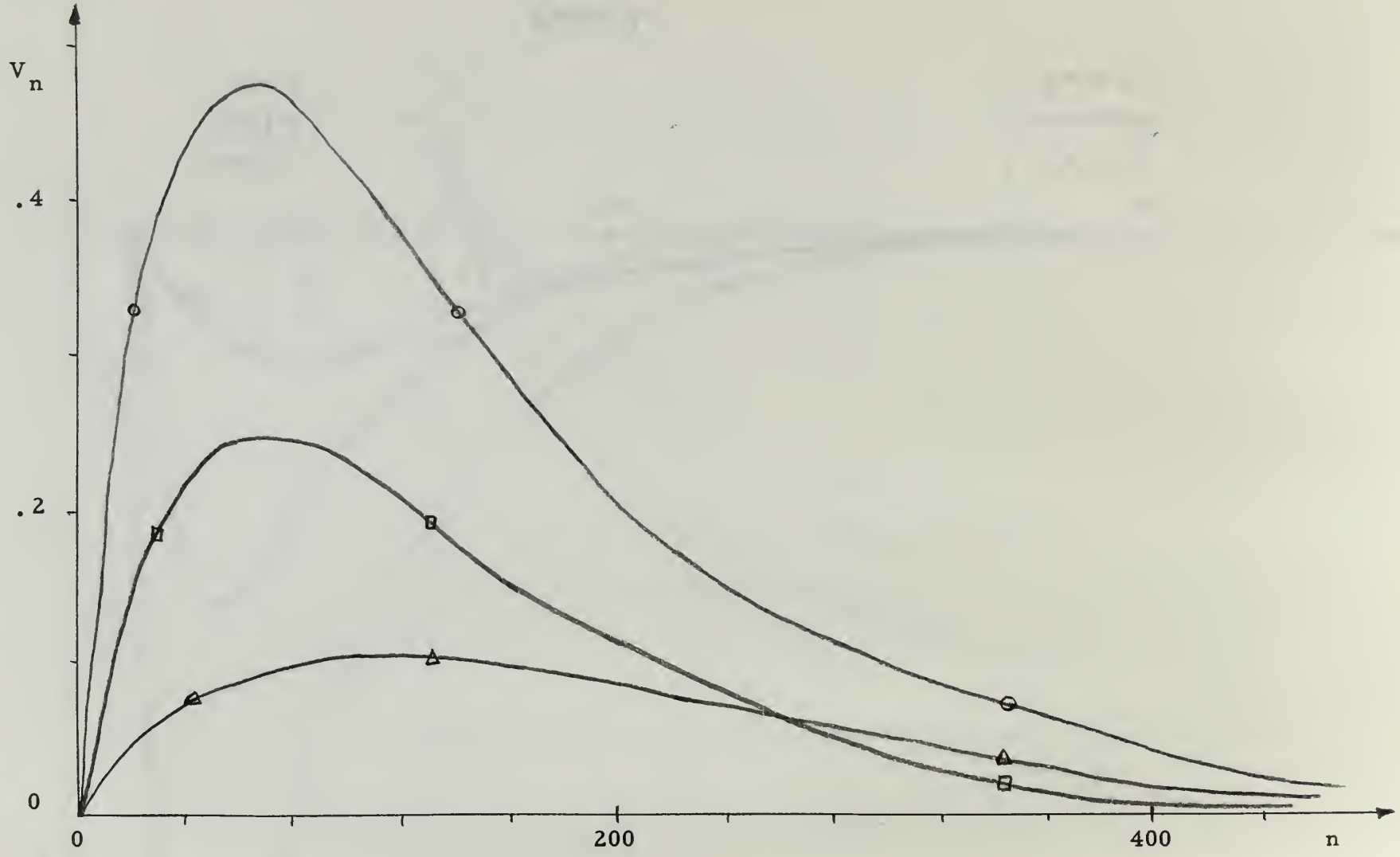


Figure 3.

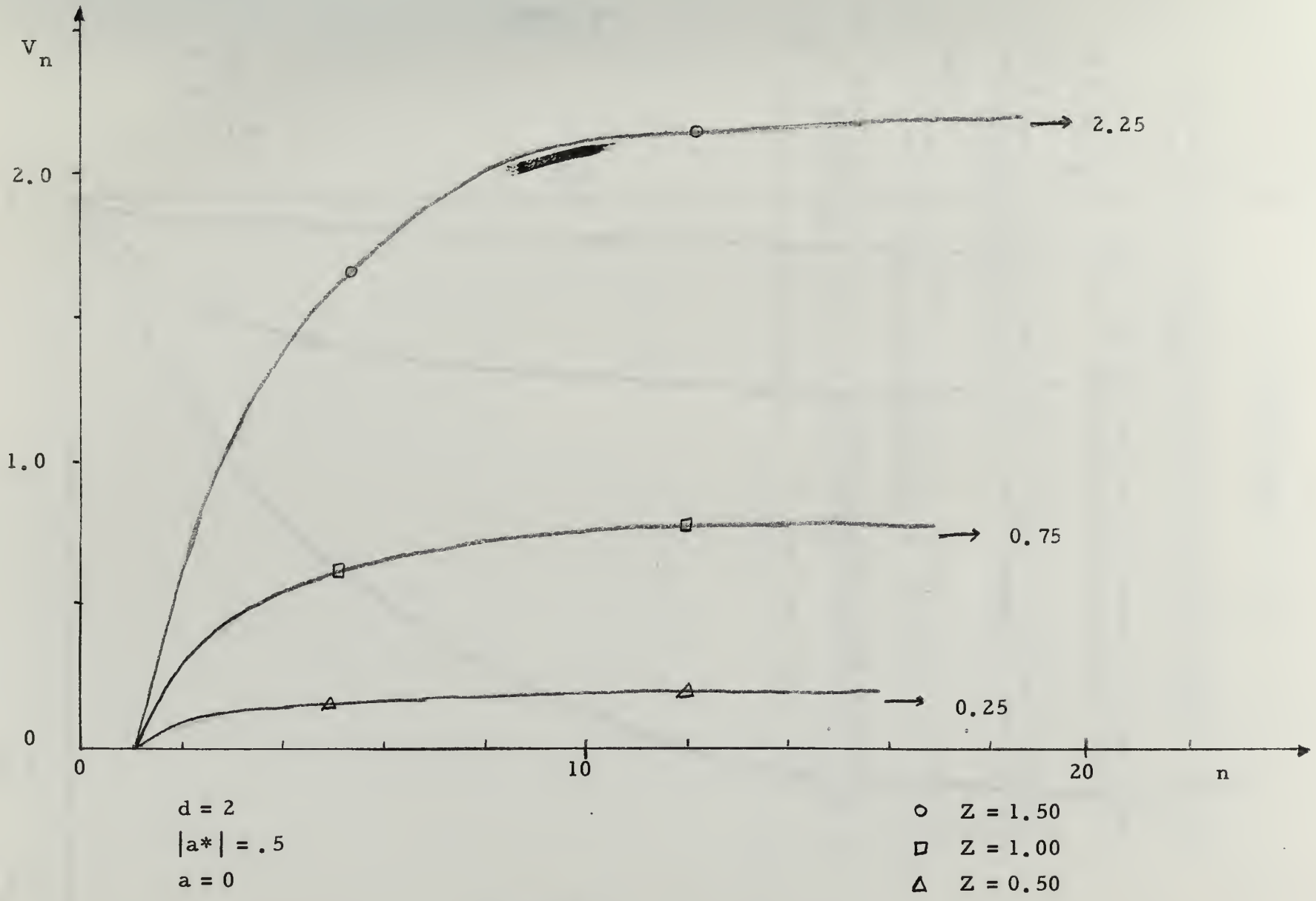
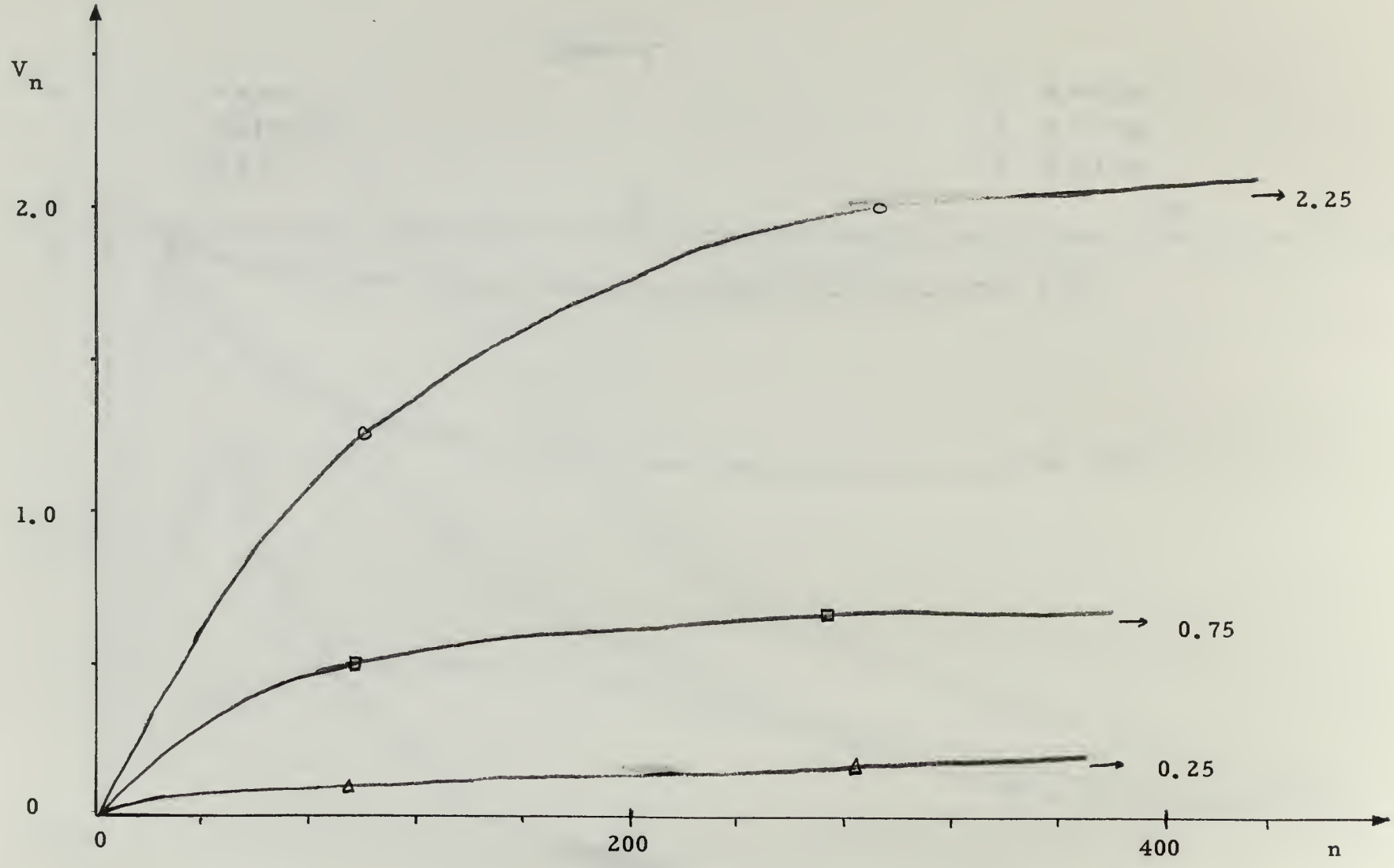


Figure 4.



d = 100
$|a^*| = .5$
a = 0

○ Z = 1.50
□ Z = 1.00
△ Z = 0.50

Figure 5.

BIBLIOGRAPHY

1. Nilsson, N.J., Learning Machines; Foundation of Trainable Pattern-Classification Systems, McGraw-Hill, New York, 1965.

2. Widrow, B., 1959. "Adaptive Sampled-Data Systems - A Statistical Theory of Adaptation," IRE Wescon Convention Record, Part 4, 1959.

3. Widrow, B. and M.E. Hoff., 1960. "Adaptive Switching Circuits." Tech. Report No. 1553-1, June, 1960. Stanford Electronics Labs. Stanford University.

4. Martinez, H.M., 1965. "A Convergence Theorem for Linear Threshold Elements", Bull. Math. Biophysics, 27, 153-159.

5. Marcus, M. and Minc, H. A Survey of Matrix Theory and Matrix Inequalities, Allyn and Bacon, Boston, 1964, Section 1.7, pf 5.

6. Halmos, P.R., Measure Theory, D. Van Nostrand, 1950. p. 194.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	20
2.	Library U.S. Naval Postgraduate School Monterey, California 93940	2
3.	Prof. Hugo Martinez Department of Mathematics U.S. Naval Postgraduate School Monterey, California 93940	1
4.	LT Robert R. Pearson, USNR 40 Woodrow Avenue Norwich, Connecticut	1

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
U.S. Naval Postgraduate School
Monterey, California 93940

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP

3. REPORT TITLE
CONVERGENCE PROPERTIES OF AN ADAPTIVE ALGORITHM FOR LINEAR THRESHOLD ELEMENTS

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Thesis, Master of Science, May 1966

5. AUTHOR(S) *(Last name, first name, initial)*
PEARSON, Robert R.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
	33	6

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	

10. AVAILABILITY/LIMITATION NOTICES
[redacted]

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT "Linear threshold element" is the generic term for a device which forms the sum $a_0 + a_1x_1 + a_2x_2 + \ldots + a_dx_d$ from an input vector $(x_1, x_2, \ldots, x_d)$ and yields one of two outputs depending on whether or not the sum is positive. A pattern classification machine may utilize a linear threshold element along with a controller which receives the one of the two values corresponding to correct classification of the input vector. The purpose of the controller is to modify the gain vector $(a_0, a_1, \ldots, a_d)$ so that the next input vector has a greater likelihood of being correctly classified by the threshold element.

This likelihood depends on the value of the gain vector and an adaptive algorithm of the "steepest descent" variety can be used to attempt to adjust the gain vector to its optimal value as the machine is exposed to a stationary sequence of statistically independent input vectors. The components of these vectors are commonly two valued, and it has been shown that convergence of the expected value of the gain vector is dependent on the value of the adjustment parameter, the values of the components, and the distribution of the input vectors. It is shown herein that a bound on the adjustment parameter, simply related to the values of the input components, is sufficient to insure this convergence. The variance of the gain vector is derived under the assumptions of a uniform input sequence and oppositely signed components of equal magnitude and it is shown that a similar bound on the adjustment parameter implies convergence of the variance. The variance is graphed under representative conditions.

DD FORM 1 JAN 64 1473

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Linear threshold element Adaptive algorithm Pattern recognition						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, syatem numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number muat be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional apace is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.

DD FORM 1 JAN 64 1473 (BACK)